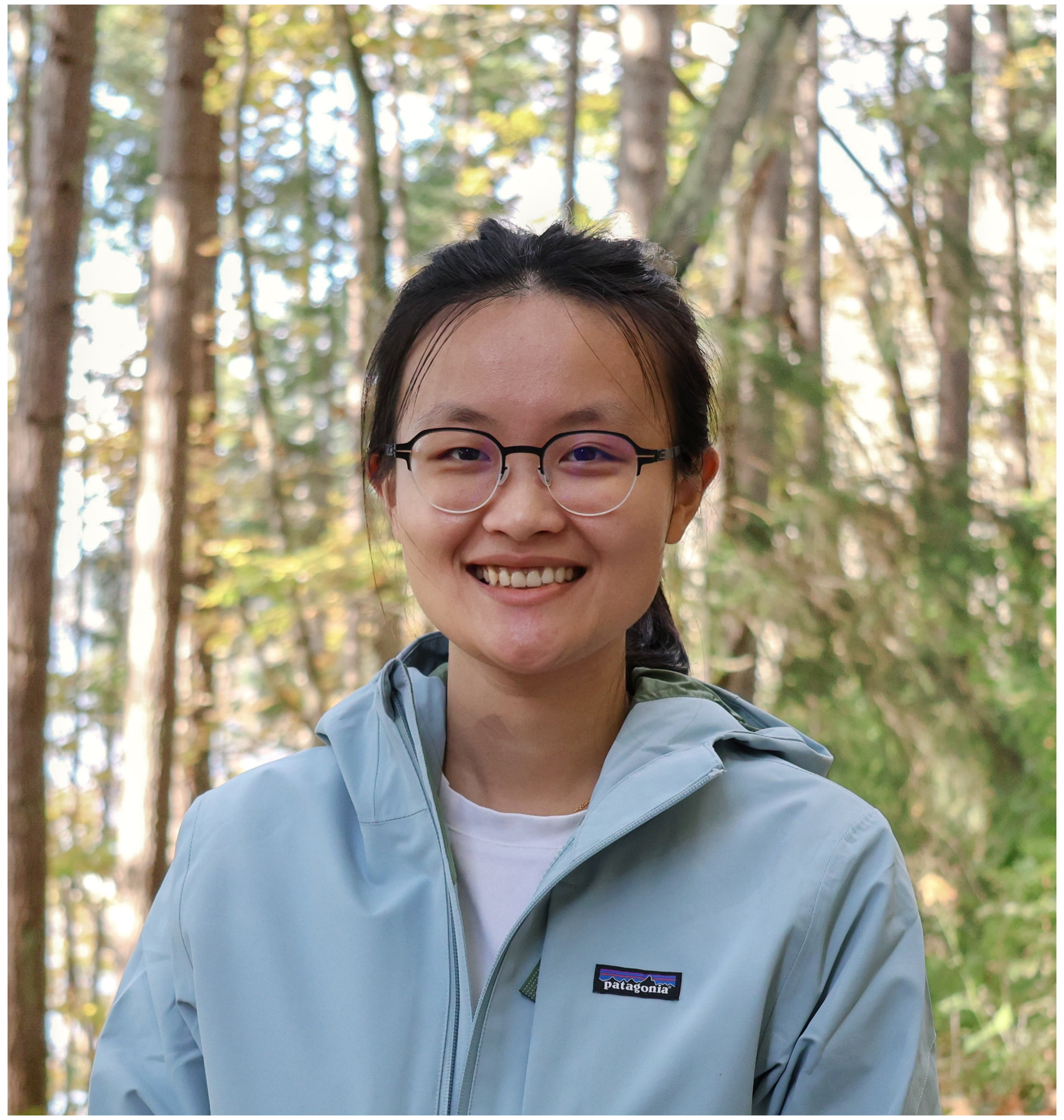
Conformal Alignment: Knowing When to Trust Foundation Models with Guarantees

Yu Gui, Ying Jin, and Zhimei Ren

The Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)

Joint work with



Ying Jin

Harvard Medical School & Data Science Initiative

Zhimei Ren

Wharton School, University of Pennsylvania

Shortage of radiologists

RSNA News

Radiology Facing a Global Shortage

Specialty affected by COVID-19, aging population and demand for imaging

BY MARY HENDERSON May 10, 2022

[Source: Radiological Society of North America]

American College of Radiology

Bulletin

Covering topics relevant to the practice of radiology

Career | Econ

How Will We Solve

As the U.S. population age

with

[Sou

Cureus. 2023 Aug; 15(8): e43866.
Published online 2023 Aug 21. doi: 10.7759/cureus.43866

PMCID: PMC10441819
PMID: 37608900

Workforce Crisis in Radiology in the UK and the Strategies to Deal With It: Is Artificial Intelligence the Saviour?

Monitoring Editor: Alexander Muacevic and John R Adler

Sadhana Kalidindi[1] and Sanjay Gandhi[2]

Author information Article notes Copyright and License information PMC Disclaimer

[Kalidindi and Gandhi, ’23]

LLM for radiology report generation

[Figure credit: MIMIC IV]

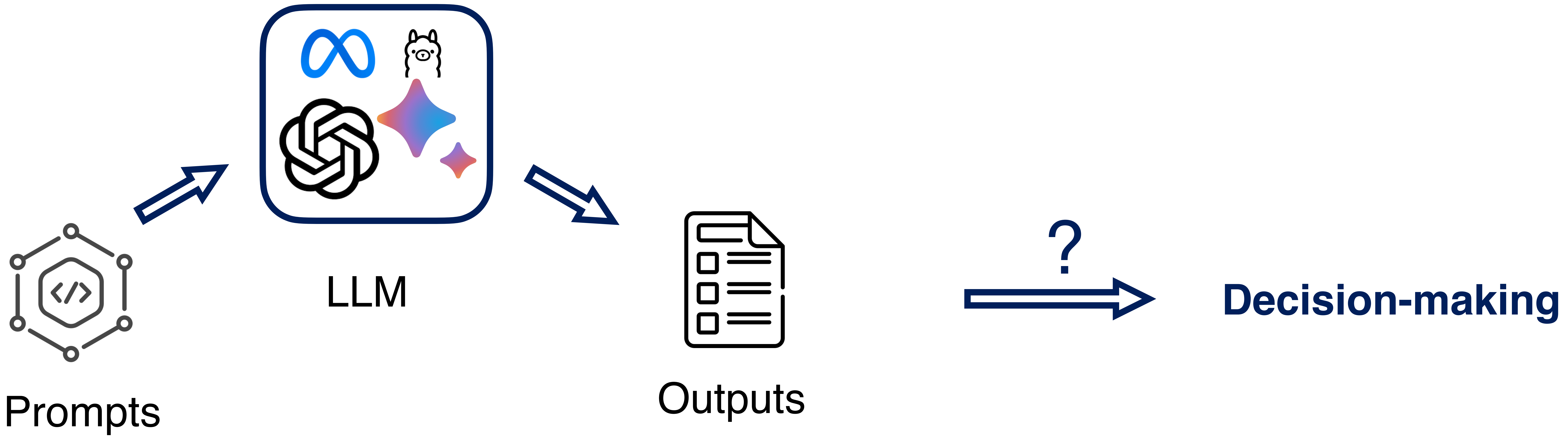
X-ray scan

In comparison with the study of _, there is little overall change. Again there is a substantial enlargement of the cardiac silhouette with diffuse bilateral pulmonary opacifications consistent with pulmonary edema. In the appropriate clinical setting, superimposed pneumonia would have to be considered.

LLM generated report

Trusted for medical decision-making?

How to safely use LLM?



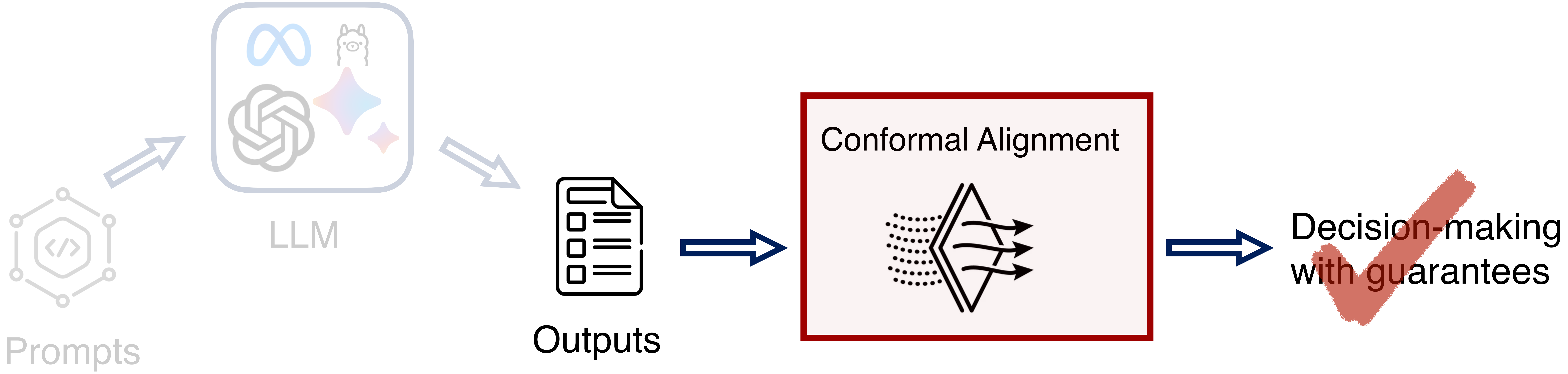
How to safely use LLM?

What guarantees are reasonable?

$\xRightarrow{?}$ **Decision-making**

How to achieve the guarantees?

How to safely use LLM?



✓Assess the alignment status

✓Identify “aligned” outputs for deployment

✓Leave the uncertain ones to experts

Evaluation of alignment

- Prompt $X \in \mathscr{X}$
  - X-ray scans, questions …
- Foundation model $f : \mathscr{X} \mapsto \mathscr{Y}$
  - Language model, vision model …
- Expert input $E \in \mathscr{E}$
  - Radiology report generated by doctors, correct answer to the question …
- Alignment function $\mathscr{A} : \mathscr{Y} \times \mathscr{E} \mapsto \mathbb{R}$

X: X-ray scan

In comparison with the study of _, there is little overall change. Again there is a substantial enlargement of the cardiac silhouette with diffuse bilateral pulmonary opacifications consistent with pulmonary edema. In the appropriate clinical setting, superimposed pneumonia would have to be considered.

$f(X)$: LLM-generated report

In comparison with the study of _, there has been a substantial increase in opacifications diffusely involving both lungs. Cardiac silhouette remains within normal limits and there is no evidence of pleural effusion. The appearance suggests diffuse pulmonary edema. However, in the appropriate clinical setting, widespread pneumonia or even ARDS could be considered.

E: expert-generated report

Chexbert [Smit et al. 04]

$$A = \mathscr{A}(f(X), E)$$

Problem formulation

- Training set: $\{(X_i, E_i)\}_{i=1}^n$
- Test set: $\{X_{n+j}\}_{j=1}^m$

Aligned units

- Wish to identify test units with $\boxed{A_{n+j} > c}$ $\iff$ testing $H_j : A_{n+j} \leq c$
- **Goal:** find a subset $\mathcal{S} \subset \{1,\ldots,m\}$ such that

$$\mathrm{FDR} = \boxed{\mathbb{E}\left[\frac{\sum_{j\in[m]} \mathbf{1}\{j \in \mathcal{S}, A_{n+j} \leq c\}}{|S|}\right]} \leq \alpha$$

"The expected fraction of selected units that are not aligned"

Predicting alignment scores

- Recall: want to select j with $A_{n+j} > c$
- But A_{n+j} is not accessible since no access to E_{n+j}
- Use predicted alignment score $\hat{A}_{n+j}$ for decision-making
- Need to account for the uncertainty of prediction to ensure FDR control

Conformal alignment

Instantiation of Conformal Selection [Jin and Candès '23]

- Divide the training data into two folds D_1 and D_2
- Model fitting: on D_1, fit a prediction model g that uses X to predict A
- Calibration: on D_2, compute the predicted alignment score $\hat{A}_i = g(X_i)$
- Conformal p-values: for each $j \in [m]$, compute the conformal p-value

$$p_j = \frac{1 + \sum_{i \in D_2} \mathbf{1}\{A_i \leq c, \hat{A}_i \geq \hat{A}_{n+j}\}}{1 + |D_2|}$$

Conformal alignment

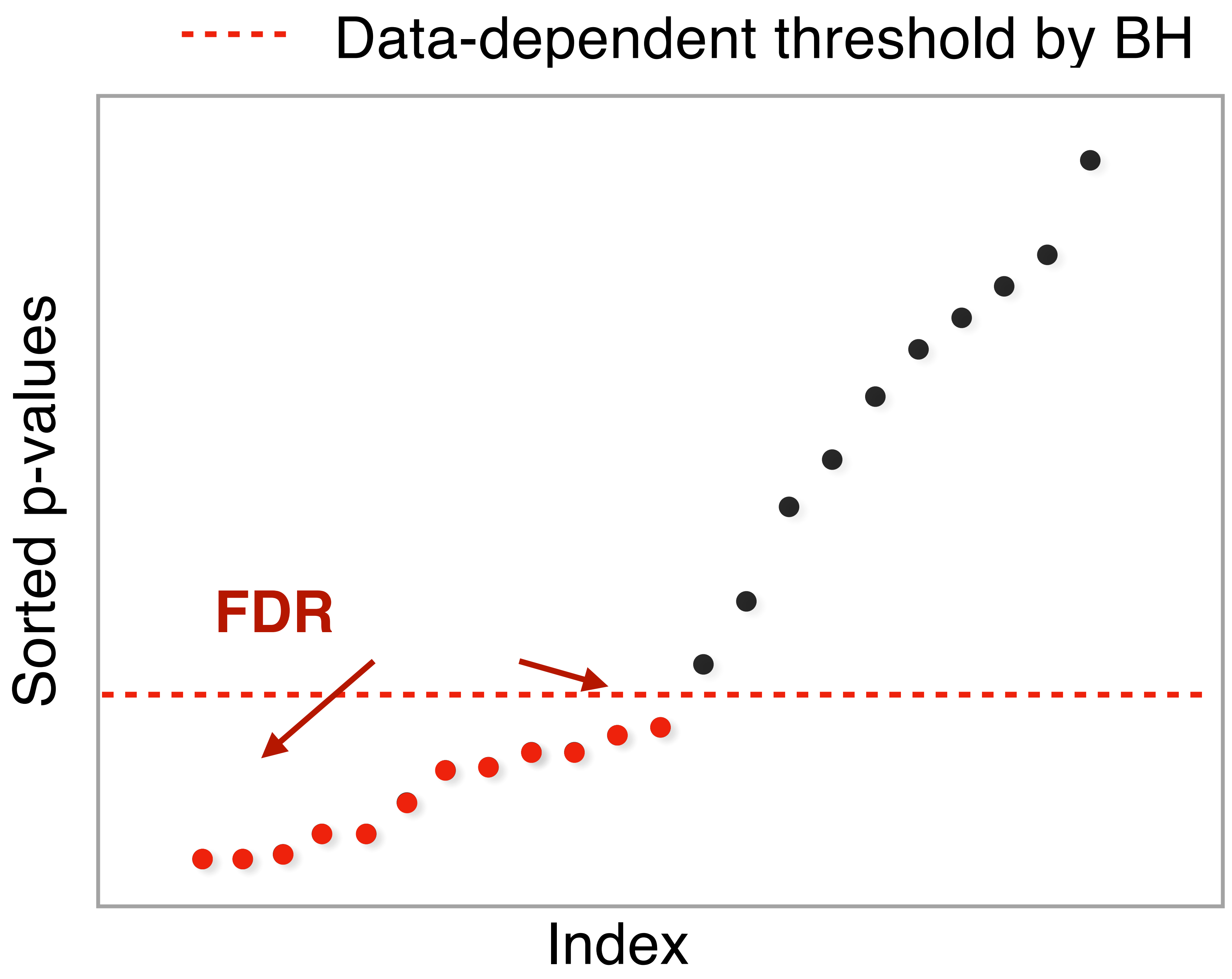
Conformal p-value $p_j = \dfrac{1 + \sum_{i \in D_2} \mathbf{1}\{A_i < c, \hat{A}_i \geq \hat{A}_{n+j}\}}{1 + |D_2|}$

Super-uniform under the null: $\mathbb{P}(A_{n+j} \leq c, p_j \leq t) \leq t$, for any $t \in (0,1)$

Selection via the Benjamini-Hochberg (BH) procedure

[Benjamini and Hochberg '95]

- Rank test samples by p-values
- Determine a "data-dependent" threshold of p-values



Theoretical guarantees

Theorem (Gui, Jin and R., 2024)

For i.i.d. data, conformal alignment at nominal level $\alpha \in (0,1)$ yields

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\sum_{j=1}^{m} \mathbf{1}\{j \in \mathcal{S}, A_{n+j} \leq c\}}{|\mathcal{S}|}\right] \leq \alpha$$

- Also applies to exchangeable data

✔ Arbitrary prediction model

✔ Arbitrary data distribution

✔ Random c

✔ Dependent data points

Desiderata for choosing g

- Evaluating the efficiency of the method

$$\text{Power} = \mathbb{E}\left[\frac{\sum_{j\in[m]} \mathbf{1}\{j \in \mathcal{S}, A_{n+j} > c\}}{\sum_{j\in[m]} \mathbf{1}\{j : A_{n+j} > c\}}\right]$$

Theorem (K., Jin and Ren, 2024)

Define $H(t) = \mathbb{P}(A \leq c, g(X) \geq t)$ and $t(\alpha) = \sup\{t : t/\mathbb{P}(H(g(X)) \leq t) \leq \alpha\}$. Under mild conditions,

$$\lim_{n,m\to\infty} \text{Power} = \mathbb{P}(H(g(X)) \leq t(\alpha) \mid A > c)$$

$$\lim_{n,m\to\infty} \frac{1}{m} \sum_{j\in[m]} \mathbf{1}\{j \in \mathcal{S}, A_{n+j} > c\} = \mathbb{P}(H(g(X)) \leq t(\alpha), A > c)$$

Desiderata for choosing g

Theorem (K., Jin and Ren, 2024)

Define $H(t) = \mathbb{P}(A \leq c, g(X) \geq t)$ and $t(\alpha) = \sup\{t : t/\mathbb{P}(H(g(X)) \leq t) \leq \alpha\}$. Under mild conditions,

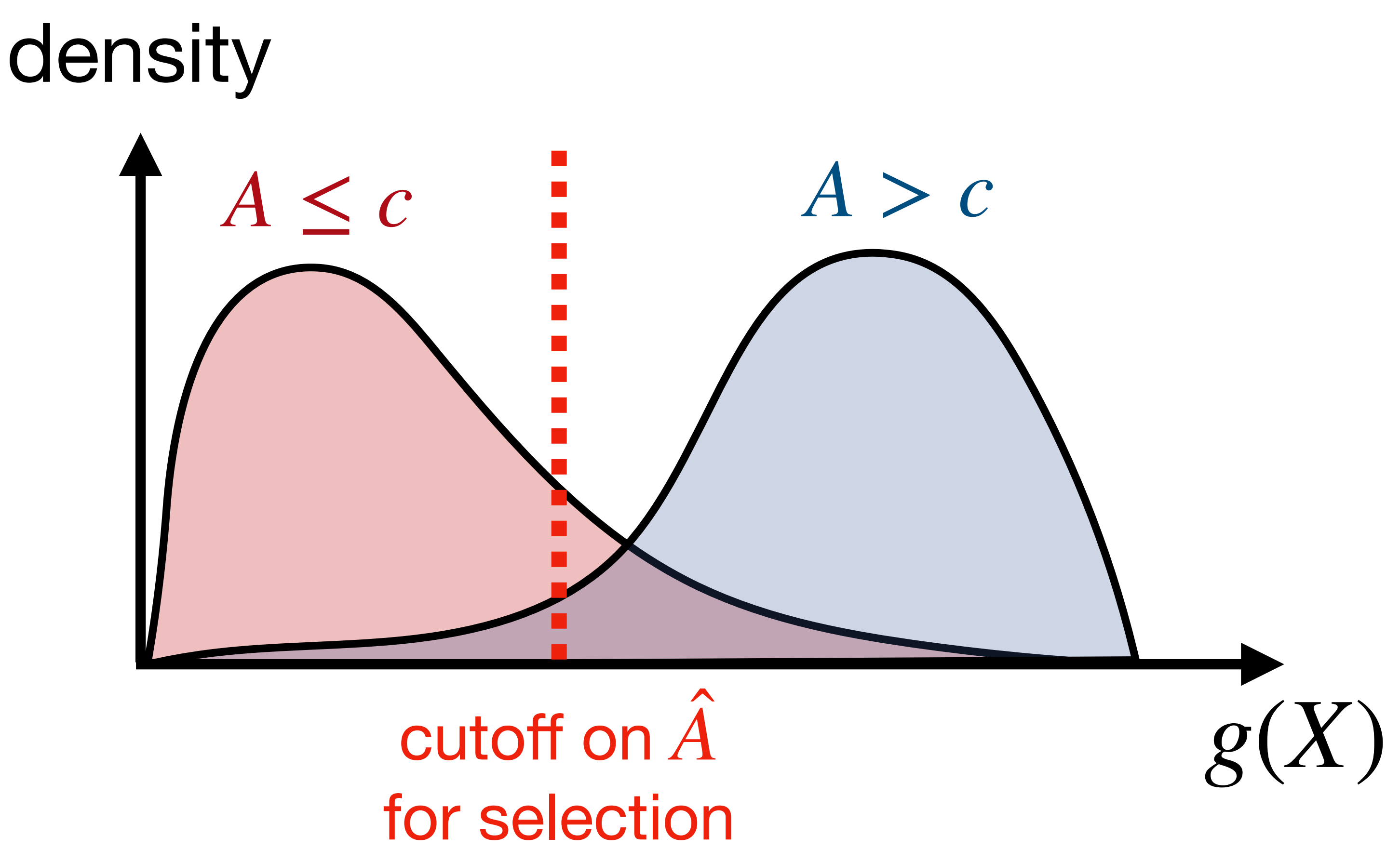
$$\lim_{n,m\to\infty} \text{Power} = \mathbb{P}(H(g(X)) \leq t(\alpha) \mid A > c)$$

$$\lim_{n,m\to\infty} \frac{1}{m} \sum_{j\in[m]} \mathbf{1}\{j \in \mathcal{S}, A_{n+j} > c\} = \mathbb{P}(H(g(X)) \leq t(\alpha), A > c)$$

The number of selections depends on

- The quality of the foundation model f (blue area)
- The quality of the prediction model g (separation)

$$\boxed{\text{Optimal } g(x) \propto \mathbb{P}(A > c \mid X = x)}$$



Example: radiology report generation

Generate report for X-ray scans

- Data: MIMIC-CXR [Johnson et al. '19]
- Prompt X: X-ray scan
- Foundation model f: fine-tuned ViT (base-patch16-224-in21k) + GPT2
- Reference E: radiology report generated by human experts
- Alignment function $\mathscr{A}$: CheXbert [Smit et al. 04]
  - convert $f(X)$ and E to two 14-dimensional vectors of binary labels
  - $A = 1$ if at least 12 coordinates match
  - $c = 0$

Predicting alignment scores

Informative & lightweight

Predictors

- Input uncertainty scores (similarity between multiple outputs)

  [Kuhn et al. ’23; Lin et al. ’23]

- Output confidence scores (functions of multiple outputs)

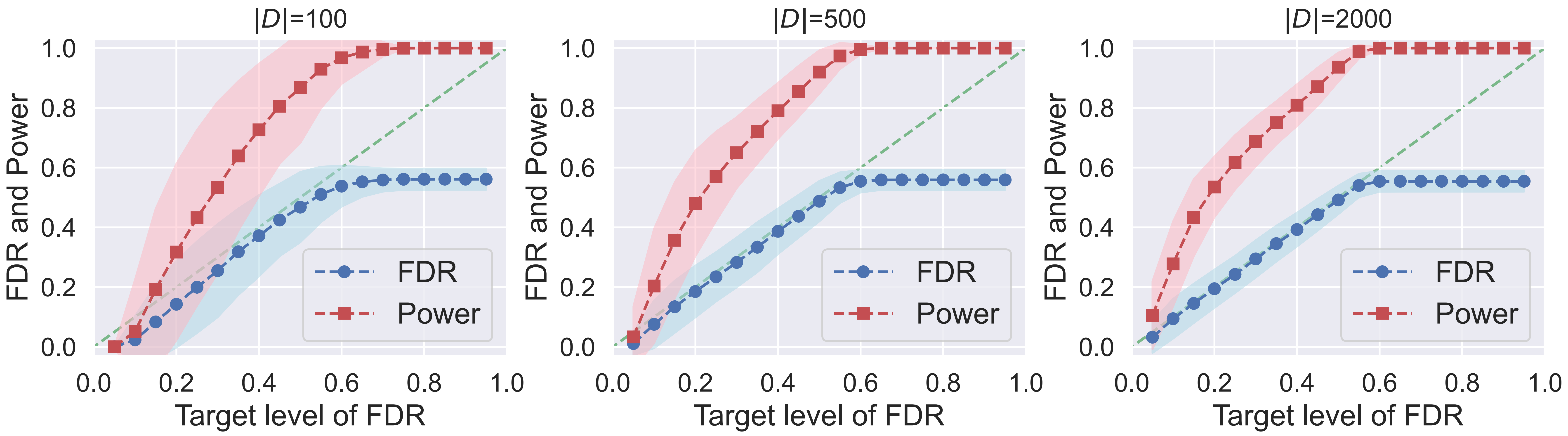
  [Lin et al. ’23]

Prediction (classification) model

- Logistic regression
- Random forest
- XGBoost

Results

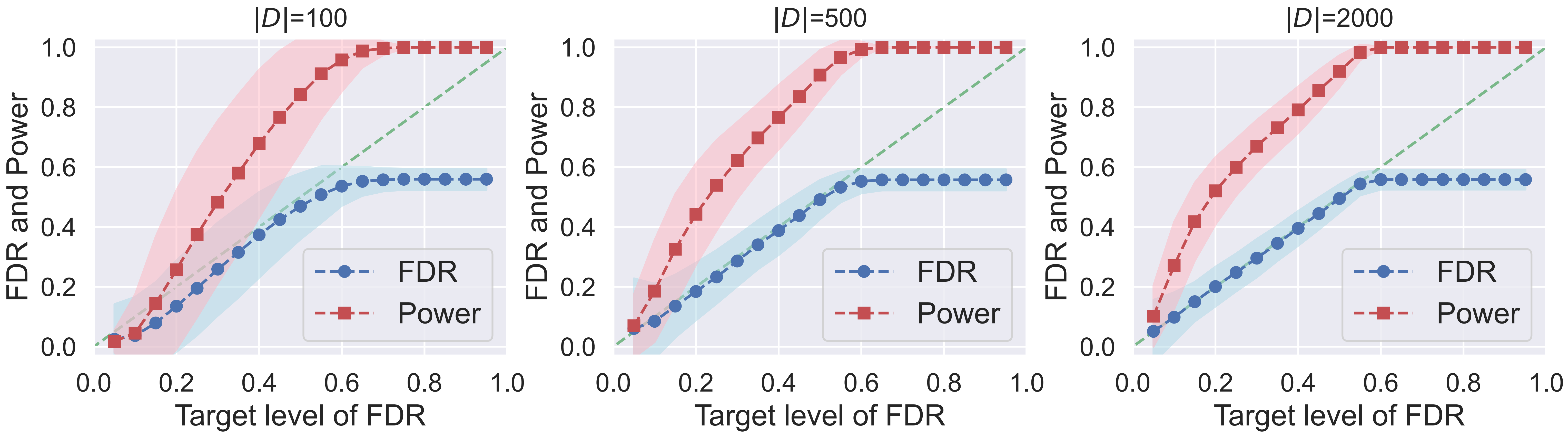
Logistic regression



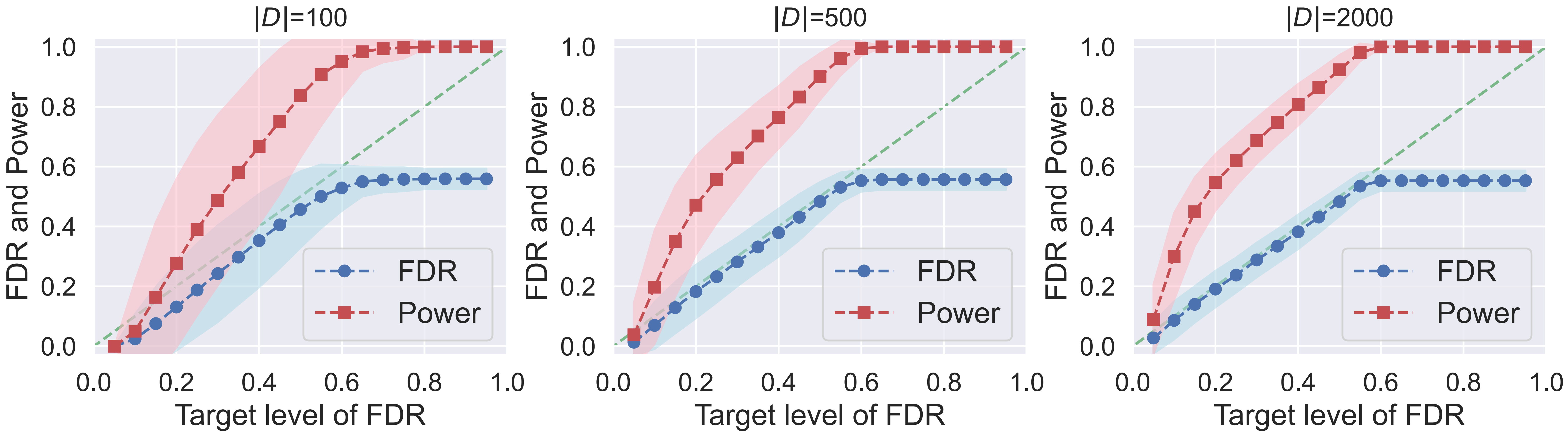
- $\gamma_1 = 0.2$ fraction of data for feature engineering
- $\gamma_2 = 0.5$ fraction of data for prediction model fitting

Effect of prediction models

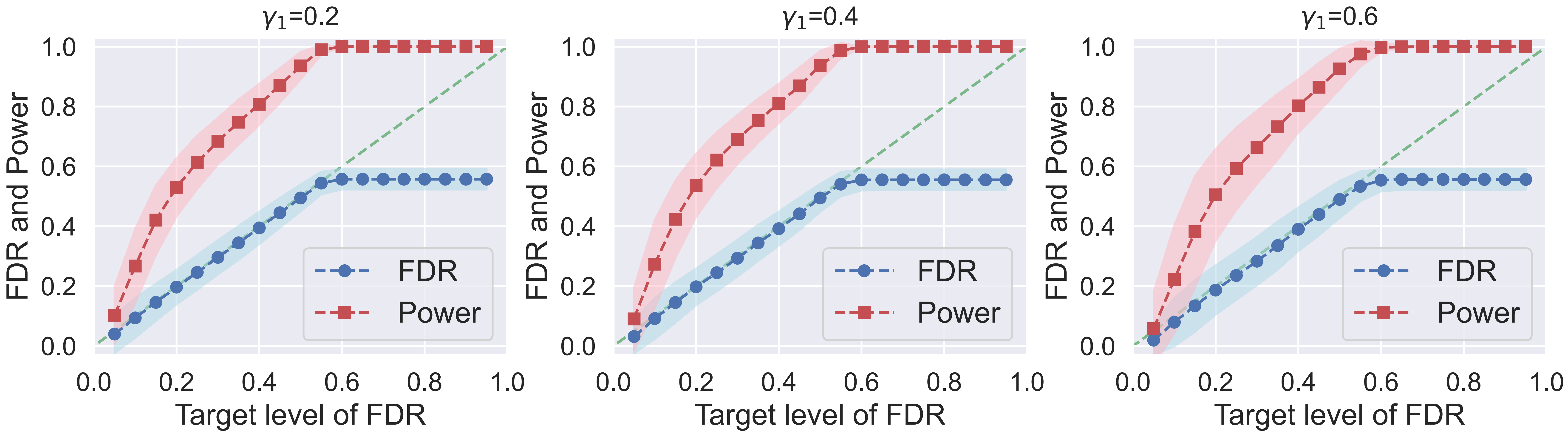
Random forest



XGboost



Effect of data partition



Example: Q&A system

More details in our paper

Conclusion

- We present Conformal Alignment that selectively deploys foundation model outputs with alignment guarantees
- The framework is instantiated in the context of question answering and radiology report generation
- Future work
  - When data arrives sequentially, can we update the model?
  - More efficient way of utilizing the referenced data

Thank you!

Gui, Y., Jin, Y., and Ren, Z. (2024). “Conformal alignment: Knowing when to trust foundation models with guarantees.” *Advances in Neural Information Processing Systems.*

https://arxiv.org/pdf/2405.10301

References

- A Smit, S Jain, P Rajpurkar, A Pareek, AY Ng, and MP Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. arxiv [cscl]. published online april 20, 2020, 2004.
- Ying Jin and Emmanuel J Candès. Selection by prediction with conformal p-values. Journal of Machine Learning Research, 24(244):1–41, 2023.
- Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1):289–300, 1995.
- Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 7:249–266, 2019.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models, 2022. URL https://arxiv. org/abs/2205.01068, 3:19–0, 2023
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023
- Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81, 2004.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221, 2022.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. arXiv preprint arXiv:2302.09664, 2023.
- Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. arXiv preprint arXiv:2305.19187, 2023.
- Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data, 6(1):317, 2019.